Tweet Coupling: A Social Media Methodology for Clustering Scientific Publications

Saeed-Ul Hassan[a], Naif R. Aljohani[b], Mudassir Shabbir[a], Umair Ali[a], Sehrish Iqbal[a], Raheem Sarwar[a], Eugenio Martínez-Cámara[c] , Sebastián Ventura[d,b], Francisco Herrera[c,b]

[a] Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan

E-mail address: saeed-ul-hassan@itu.edu.pk; mudassir.shabbir@itu.edu.pk; mscs15013@itu.edu.pk; sehrishiqbal@itu.edu.pk; raheem.bwl@gmail.com

[b] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia E-mail address: nraljohani@kau.edu.sa

[c] Andalusian Research Institute in Data Science and Computational Intelligence **(**DaSCI**)**, University of Granada, 18071 - Granada, Spain E-mail: herrera@decsai.ugr.es; emcamara@decsai.ugr.es

[d] Andalusian Research Institute in Data Science and Computational Intelligence **(**DaSCI**)**, University of Córdoba, 14071 - Córdoba, Spain E-mail: sventura@uco.es

Abstract:
We argue that classic citation-based scientific document clustering approaches, like co-citation or bibliographic coupling, lack to leverage the social-usage of the scientific literature originate through online information dissemination platforms, such as Twitter. In this paper, we present the methodology *tweet coupling*, which measures the similarity between two or more scientific documents if one or more Twitter users mention them in the tweet(s). We evaluate our proposal on an altmetric dataset, which consists of 3,081 scientific documents and 8,299 unique Twitter users. By employing the clustering approaches of bibliographic coupling and tweet coupling, we find the relationship between the bibliographic and tweet coupled scientific documents. Further, using VOSviewer, we empirically show that tweet coupling appears to be a better clustering methodology to generate cohesive clusters since it groups similar documents from the subfields of the selected field, in contrast to the bibliographic coupling approach that groups cross-disciplinary documents in the same cluster.

Keywords: Scientific Document Clustering, Social Media, Altmetrics, Tweet Coupling, Bibliography Coupling,

1. Introduction

Clustering scientific documents aims to organise the set of documents into groups, such that documents in a single group are similar to each other in comparison to the documents in other groups (Lawrence, Bollacker, & Giles, 1999; Thijs and Glänzel, 2018). The clustering of scientific documents is crucial for several tasks, such as summarisation (Karimi et al., 2018), recommendation systems (Habib and Afzal, 2019), semantic understanding of scientific research (Shardlow et al., 2018), classification of scientific documents (Heffernan, K., & Teufel, 2018), and information retrieval systems for digital libraries (Safder & Hassan, 2019). However, the clustering of related scientific documents in growing scholar big data is a challenging task (Hassan and Haddawy, 2013; 2015). There are several known classic approaches to cluster similar scientific documents such as bibliographic coupling (Martyn, 1964), co-citation (Small, 1973) or Amsler (1972) approach. These existing approaches cluster similar scientific documents using the meta-data of the scientific documents' references, venues, authors, keywords among other features.

The limitations of the classic approaches are two-fold:

1. They do not leverage the user perspective on the scientific literature. As a result, the most relevant documents against a cluster are often missed out, that actually best match in accordance to users' perception (Mesbah et. al., 2017).
2. The classic citation-based methods come along with the inherent issue of publication and citation time lags.

We claim that these limitations can be addressed by clustering the publications based on the real-time usage of scientific publications or discussion of scientific literature on social media platforms. People are increasingly going online to find and share the information about science. Specifically, researchers are using the social media platforms to engage with each other. Altmetrics offers innovative tools for researchers to explore the public engagement with science in social media platforms. Consequently, new possibilities are emerging to analyse the interaction between researchers and research articles on social media platforms (Hellsten & Leydesdorff, 2017, Hellsten et. al., 2019, Joubert & Costas, 2019, Robinson-Garcia et. Al., 2019).

In order to address the previous drawbacks, in this paper, we present tweet coupling, which is a new methodology to measure the similarity of documents by leveraging the social usage of scientific documents on Twitter platform. The main advantage of tapping user engagements pertaining to the scientific publications on social media plateforms is that they are much faster than citation counts, which at least take a few years after the publication of an article to be ready for the evaluation purpose (Costas et al., 2015; Haustein et al., 2015; Shu et al., 2018, Ananiadou et al., 2013).

Tweet coupling is similar to classic bibliographic coupling approach. According to Martyn (1964), two scientific papers are bibliographically coupled if they have at least one common reference. If paper *A* and *B* refer paper *C*, it indicates a potential relationship between paper *A* and *B*, therefore, paper *A* and *B* are said to be bibliographically coupled. Thus, documents would have more coupling strength if they have a large number of common references. Similarly, tweet coupling is defined as follows: If a Twitter user mentions paper *A* and *B* in either same or two different tweets, then we assume this reflects a relationship between the papers and we called the papers as 'tweet coupled'. In other words, two papers are tweet coupled if they have at least one common Twitter user. Thus, with a large number of common Twitter users reflect a high 'tweet coupling' strength. Formally, 'tweet coupling' can be described as follows: Let $U = \{u_1, u_2, u_3 \cdots u_n\}$ be the set of Twitter users, $T = \{t_1, t_2, t_3 \cdots t_n\}$ be the set of tweet text by tweet users, and $D = \{d_1, d_2, d_3 \cdots d_n\}$ be the set of scientific documents mentioned in T_i by U_i. Let $D_{u_i} = \{d_{u_1}, d_{u_2}, d_{u_3} \cdots d_{u_n}\}$ be the set of documents that a given user u_i mentions in tweet t_i. Formally, two set of documents are tweet coupled *iff* $D_{u_i} \cap D_{u'_j} \neq \emptyset$ and $u \neq u'$.

Since our employed solution relies on analysing user engagements on scientific documents under the umbrella of altmetrics, we briefly describe the phenomenon of altmetrics in the context clustering similar scientific documents (see Section 2.2 for detailed discussion). Altmetrics term was introduced in 2010 by "Jason Priem" as an abstraction of social web metrics (Priem, 2010). Nowadays, altmetrics[1] becomes a novel source to measure the social activities regarding scientific literature as well as it provides futuristic metrics which complement conventional bibliometric that solely depend on the citation counts, number of publications and peer review (Butler et al., 2017).

[1] https://www.altmetric.com/

Altmetrics uses various social media as a data source such as *Twitter*, *Facebook*, *Google+*, *Linked-in*, etc. It tracks all relevant event such as *like*, *comment*, *share*, and *retweet* on any research article which gives us usage metrics of that article (Priem & Costello, 2010; Haustein et al., 2015; Zahedi et al., 2014; Hassan et al., 2017; Said et al., 2019). As mentioned earlier, the major advantage of altmetrics is that they are much faster than citation counts which at least take a few years after the publication of an article, to be ready for the evaluation purpose (Costas et al., 2015; Haustein et al., 2015; Shu et al., 2018).

Recently, Twitter has received significant attention with plenty of opinions about scientific documents. Specifically, researchers share their work on Twitter, discuss modern topics and talk about the research informally by commenting, liking and retweeting on certain posts (Adie & Roe, 2013; Thelwall et al., 2013). Note that among all the altmetric platforms, Twitter has the highest coverage i.e. 87.1% (Robinson-García, et al., 2014, Robinson-García et al., 2017). Thus, it makes Twitter a significant and well-suited platform to obtain user engagement statistics, but any other social media platform could be used to conduct this investigation, e.g., Mendeley. To conduct experiments, we utilize the dataset of scientific documents from the field of Library and Information Sciences from Scopus. At first, we cluster the scientific documents using bibliographic coupling and tweet coupling, respectively. Further, we find similarity between bibliographic and tweet coupled document. Next, we visualize and compare the relationship of bibliographic and tweet coupling using VOSviewer. Finally, we discuss the implication of our employed tweet coupling measure and its applications for the scientific document search applications such as classification of scientific documents, recommendation systems, and information retrieval systems for digital libraries.

The contributions of this paper are:

- The description of tweet coupling, which is a new methodology to measure the similarity of documents by leveraging the social usage of scientific documents on Twitter platform.
- The study of the relation among tweet coupling and traditional citation-based metrics.

The rest of the paper is organized as follows: Section 2 describes the detailed literature review, including existing coupling techniques for document clustering. Section 3 presents our method for collecting data, employed tweet coupling approach for document clustering and similarities

between tweet coupling and bibliographic coupling. In section 4 we present the result of our experiments and detailed comparison between bibliographic coupling, and tweet coupling. Finally, Section 5 presents some concluding remarks and indicate future directions of this research.

2 Background

In this section, we review the relevant literature on bibliographic coupling in Section 2.1, the use of altmetrics data in bibliographic studies in Section 2.2, and other works related to our proposal.

2.1 A brief review on bibliographic coupling

The practicality and success of scientific work are often measured by the attraction it receives from the scientific community as well as the quantitative measure of the scientific work that extends it (Garfield, 1979, Batista-Navarro et al., 2013). In order to find related work, there are different approaches exist to determine the similarity of scientific documents. Most of the time, citation analysis gives excellent result to find document similarity. There are number of citation analysis techniques that are used for the identification of similar scientific documents. Amongst them, co-citation, bibliographic coupling, citation proximity, and Amsler method are the most widely and easily applicable citation techniques however, each one has their own pros and cons. In co-citation, two documents are co-cited if both document cited by at least one paper in common (Small, 1973). For example, paper A and paper B are co-cited if both A and B paper appear in the references of third paper (Gipp & Beel, 2009). It is used to find out the semantic similarity between research publications. If two papers received more co-citation, there citation strength is higher, and they are more likely to be semantically relevant as well. Co-citation is a forward-looking assessment technique. The drawback of this technique is that if the paper is recently published and it has no citation, so it is hard to find out the semantic relationship with other papers using co-citation. This technique is useful for those papers only which have a high citation rate.

It is in contrast to co-citation, two documents are bibliographically coupled if they are sharing at least one common reference in a bibliography (Kessler, 1963). For example, paper A and B are bibliographically coupled if paper C is in the bibliography of both A and B (Gipp & Beel, 2009). Similar to co-citation, a number of studies have used bibliographic coupling as a measure of semantic similarity between the scientific documents (Trueger et al., 2015; Zhao & Strotmann, 2014). If there are large number of common references in papers, their bibliography strength is

high and they are more likely semantic related. Bibliographic coupling is backward-looking assessment technique. The advantage of this method is that we can also find a semantic relationship of newly published papers with others.

Amsler, (1972) proposed a measure of similarity between two documents that combine both co-citation and bibliographic coupling. According to Amsler, two papers A and B are related if A and B are cited by the same paper, A and B cite to the same paper. Let d is the document and P_d is the set of parents (cite papers) of P and C_d is the set of children (citatuons) of d. The Amsler similarity between two documents measures as shown in Eq. 1:

$$Amsler(D_1, D_2) = \frac{(P_{D1} \cup C_{D1}) \cap (P_{D2} \cup C_{D2})}{|(P_{D1} \cup C_{D1}) \cup (P_{D2} \cup C_{D2})|} \quad (1)$$

Citation proximity analysis is the enhancement of co-citation analysis, consider the proximity of citation to each other within an article full-text (Gipp & Beel, 2009). Citation proximity index can be (CPI) calculated in three steps. In the first step, documents are parsed and position of citation in the document is analysed. In the second step, each citation is assigned to the corresponding items in the bibliography. In the last step, the proximity between each pair of citation is analysed, if they are closer to each other than there are more chances that they are related to each other. For example, two citations are given in the same sentence their CPI is 1 as if they are in the same paragraph, CPI is 1/2. If it is in the same chapter, CPI is 1/4.

Yan & Ding, (2012) explored the similarity between six types of scholarly network including co-citation network, bibliographic coupling network, co-authorship network, co-word networks and topical networks. Cosine distance was chosen to find the similarity between all these networks. They found that citation network and co-citation network; bibliographic coupling network and co-citation network; and co-word networks and topical networks have high similarity whereas, topical network and co-authorship network have low similarity. They recommended using hybrid network to analyze research interaction and scholarly communication. Since this investigation relies on the use of user perception of scientific literature on social media, the following subsection reviews on existing almetric studies in the context of clustering scientific documents.

2.2 A Brief Review on Altmetric studies and Social Network Analysis

Citation counts are frequently used for the evaluation of scientific research. However, the disadvantage of using citation counts to evaluate the scientific research is that they are quite slow. Altmetrics is an alternative indicator which is derived from social media and provide quicker scientific impact (Mohammadi & Thelwall, 2014, Nawaz et al., 2012). People are increasingly going online to find and share the information about science (Hellsten & Leydesdorff, 2017, Hellsten et. al., 2019, Joubert & Costas, 2019, Robinson-Garcia et. Al., 2019). Specifically, the researchers have been urged to consider how they can use the social media platforms to engage with each other. Altmetrics offers innovative tools for researchers to explore the public engagement with science in social media platforms. Consequently, new possibilities are emerging to analyse the interaction between researchers and research articles on social media platforms.

Several studies can be found that are focused on socio-semantic analysis of the scientific publications (Hellsten & Leydesdorff, 2017, Hellsten et. al., 2019, Joubert & Costas, 2019, Robinson-Garcia et. Al., 2019). Joubert & Costas (2019) conducted an investigation to expand the understanding of the relationships and interactions between social media users and scientific outputs. They explored the identities, characteristics and activities of South African science tweeters—i.e. Twitter users in South Africa who tweet about research articles. The growing number of science tweeters, both overall and in relative terms, suggests that Twitter users are increasingly using this social media platform as a tool to share and discuss scientific outputs. The science tweeters are actively contributing to the sharing of information about new research articles. Moreover, several studies can be found that focused on identifying the topics of interests and the communities of users using altmetrics data (Hellsten & Leydesdorff, 2017, Hellsten et. al., 2019, Joubert & Costas, 2019, Robinson-Garcia et. Al., 2019). For example, Robinson-Garcia et. Al., (2019) identified the topics of interest within the field of Microbiology and identify the main sources driving such attention. Specifically, they combined the data from Web of Science and altmetric.com to conduct their investigation. They found that a central area of the network is formed by papers discussed by the three outlets. Their topic analysis shows that the thematic focus of papers mentioned varies by outlet.

The application of altmetrics and social networks are expanding significantly, however, the novelty of this work is the usage of altmetrics for the clustering of scientific documents. To the best of our knowledge, no attempt has been made to use social media contents such as tweets on scientific publications as a proxy to measure the similarity among the papers.

Among all the altmetrics data sources, Twitter is the most widely used platforms by the scientific community. Priem et al. (2010) investigated 46,515 tweets from the sample of 28 scholars and concluded that Twitter citations are much faster as compared to traditional citation measures (Melero, 2015). In addition, Priem et al. (2012), analyzed the correlation of altmetrics with citation count and showed that there exists a significant contribution of altmetrics in citation prediction of research. An analysis across more than 40 cross metric validation studies presented a weak correlation between citation count and altmetrics ranging from 0.08% to 0.5% (Erdt, Nagarajan, Sin, & Theng, 2016).

Hassan and Gillani (2016), measured the impact of the altmetric field. They collected data from social media sites including Twitter, Facebook, Mendeley, CiteUlike and Wikipedia for the years 2010 to 2014. The information gathered was only related to authors working in the field of altmetric. All scholarly information is gathered from Google Scholars database. Dataset consists of relevant information on a total of 47 distinct scholars. They introduced alt-index similar to h-index, based on altmetric count of the scholars. They observed that Pearson's correlation of $\rho = 0.247$ between h-index and alt-index. A relatively high correlation was observed between social citation and scholarly citation with $\rho = 0.646$. Moreover, Peoples, Midway, Sackett, Lynch, & Cooney, (2016) find the relationship between traditional metrics of research impact and modern altmetrics specifically twitter activities to measure the research impact of a research article. They used the dataset of 1,599 research article from 20 ecology journal published from 2012- 2014 and found a strong positive correlation between citation count and unique tweet count on research publications. According to them, twitter activities were not dependent on the impact factor of journal, the highest impact journals were not compulsory the most tweets on twitter. Their results concluded that altmetrics and traditional metrics can be useful to find research impact and closely similar to each other but not exactly the same.

Liu & Fang, (2017) investigated 79,441 English written tweets of top 100 research article published in 2015. They categorized the tweet among different categories and recommended that tweet written by those involved in the publication of paper should not be considered to measure the impact of the research article. They proposed to omit the tweets with the context that is irrelevant to the paper and tweets with a negative opinion should also be omitted. Tweets with positive sentiments and neutral tweets which also represent agreement towards paper to a certain degree should be considered only while evaluating twitter impact. After analyzing the tweet text, comprehensive list of positive and negative words or phrases were presented that are majorly used among researcher, while sharing their opinion about research work. They verify its correctness by searching these terms in a large data set of tweets. These words were then also added in SentiStrength lexicons (Thelwall et al., 2013). More recently, Didegah and Thelwal (2018) presented a comparative study by investigating network level differences between citations, Mendeley saves, and tweets for research articles. They surprisingly found minor overlap between these three phenomena.

Older publications have lower coverage of altmetrics scores due to the less prevalent use of social web at the time of publication. Comparatively, more recent research publications have much higher altmetrics counts (Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013, Haustein et. al., 2016). Additionally, Holmberg & Thelwall, (2014) examined the cross-disciplinary usage of twitter, how and why they use twitter and to see whether there exist a common pattern of usage among different fields. Different discipline(s) tweets were analyzed and categorized in different groups. Their result showed that a clear difference in twitter usage among scholars in these disciplines. Zahedi, Costas, Larivière, & Haustein (2017) examined the characteristics of scientific literature and types of people that share and discuss their research work on social media. Dataset on which they worked contained 1.3 million records having combined, both scholarly and social information. After that different document features (*document type*, *number of pages*, *cited sources*, *characters in the title*, *number of authors*, *countries of origin*, and *affiliated institutions)* were computed. Based on their result, Social media coverage is very low, with 22.6% of papers receiving at least one tweet, 5.2% publically shared on Facebook, 2.3% mentioned in a blog post and 1.1% discussed by mainstream media (Zahedi et al., 2014).

2.3 Summary and Comparison with our Work

The literature review presents an array of studies that use citations among scientific publications to determine their semantic relatedness. As discussed earlier in Section 1, the limitations of the bibliographic coupling techniques are twofold: a) These methods do not leverage the user engagements on the scientific documents. As a result, most relevant documents against a cluster are often missed out, that actually best match in accordance to users' perception. b) The classic citation-based methods come along with the inherent issue of publication and citation time lags.

In this paper we introduce the methodology tweet coupling which is built upon a methodology for clustering scientific publications according to their real-time usage on Twitter. One of the main advantages of exploiting user engagements of scientific publications on Twitter plateform is that they are much faster than citation counts which at least take a few years after the publication of an article, to be ready for the evaluation purpose (Costas et al., 2015; Haustein et al., 2015; Shu et al., 2018, Ananiadou et al., 2013). To the best of our knowledge, no attempt has been made to use social media contents such as tweets on scientific publications as a proxy to measure the similarity among the papers. Next section elaborates the employed measure of tweet coupling and compares it with conventional bibliography coupling.

3 The Tweet Coupling Methodology

In this section, we describe the tweet coupling methodology that is depicted in Figure 1. The methodology is composed of two steps, which are the building of a coupling incidence matrix described in Section 3.1, and the building of the adjacency matrix from the incidence matrices detailed in Section 3.2.

The application of the methodology begins with the identification of the scientific papers which are tweet coupled and bibliographically coupled using altmetric and Scopus reference list respectively. Subsequently, we identify a reference list of all 1537 papers from the Scopus database to compute bibliographically coupled papers. Next, we tap the social activities of these papers on twitter platform using the altmetric database to compute tweet coupled papers. Finally, we measure the Jaccard similarity between bibliographically and tweet coupled papers to study their relationship.

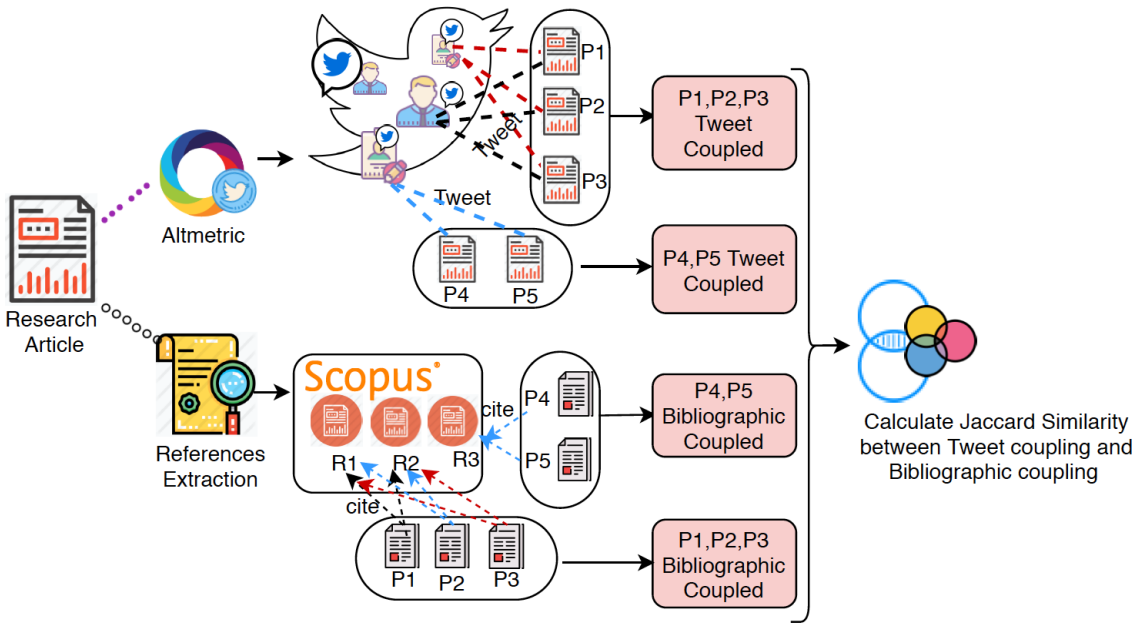

Figure 1: Flow diagram of data inputs and processing

3.1 Coupling Incidence Matrices

In order to compute bibliographic coupling, we generate an incidence matrix between scientific papers and their references. Similarly, to compute tweet coupling we generate incidence matrix between scientific papers and twitter users. Incidence matrix gives the relation between two classes of objects. One class along the rows of matrix (i.e. scientific papers) other class along the column (i.e. references or twitter users). Each row represents the single research article and each column represent the single reference. If a reference occurs in the bibliography of a given paper then the intersection of row (scientific paper) and column (reference or twitter user) is placed with '1' while we placed '0' on the intersection of row (scientific paper) and column (reference or twitter user) otherwise.

3.2 Bibliography and tweet coupling

In the next step, we compute adjacency matrices from incidence matrices which give us bibliographic and tweet coupling matrices. An adjacency matrix is a square matrix which gives us the connection between two objects of the same class. In the case of a graph adjacency matrix,

rows and column are labeled with graph vertices in the matrix and their intersection represents the connection or an edge between these two vertices. The diagonal of the adjacency matrix is traditionally labeled as 0, for a simple graph. We will construct adjacency matrices from the relevant incidence matrices defined above i.e. 1. The square matrix is defined in Eq. 2.

$$A_{squareMatrix} = B * B^T \tag{2}$$

Entries in matrix *A* represents the relation between a pair of scientific papers. The value represents the status of the connection, if the value is 0 on intersection its means that these two scientific papers are not bibliographically or tweet coupled. If the value is greater than 0 its means these two papers are bibliographically or tweet coupled. Larger intersection value signifies strong semantic relation between the scientific papers. In our square matrix, diagonal values represent the total references or twitter users on each scientific paper.

Further, in order to measure a meaningful correlation between bibliographic and tweet coupling square matrices, we convert the incidence matrices to binary matrices by replacing all non-zero values of square matrices with 1. Furthermore, we also connect all those papers which are directly not connected but indirectly connected via any other paper in both tweet coupled and bibliographically coupled matrices. Using the Jaccard measure, we calculate the similarity between the two matrices. The Jaccard measures similarity score by taking a ratio between a common and distinct member of the tweet and bibliographic coupling matrix. Given two scientific papers P_1 and P_2 , their Jaccard similarity can be computed as shown in Eq. 3.

$$SIM_j = \frac{P_1 \cup P_2}{P_1 \cap P_2} \tag{3}$$

The Jaccard similarity coefficient ranges from 0 to 1. It is 1 when P_1 and P_2 are similar to each other and 0 when they are completely different (Huang, 2008).

4. Results and discussion

This section presents the dataset (see Section 4.1), evaluation measures (see Section 4.2), and the comparison of the results among Tweet coupling and bibliographic coupling (see Section 4.3) between bibliographic coupling and tweet coupling.

4.1 Data and pre-processing

The data used in the experimentation was given by altmetric.com, on June 14, 2016. There was a total of 4.5 million JSON files in the dataset. Each file contains information about the single article and respective articles can be identified uniquely by an altmetric id. Our dataset contains all altmetric data from July 2011 to June 2016 and there was a total of 3081 scientific publications. From this initial dataset, we filtered out the publications that belong to the Library and Information Sciences Journals, using All Science Journal Classification adopted by Scopus. Since altmetric data provides information about the online web indices, so references were collected from Scopus using Scopus API by using article DOI (or article title in cases where DOI's were not available). To get the tweet details, we used the tweet-id which is given in altmetric data for every 3081 publications. We used twitter API to fetch details of each tweet such as *tweet text, name, screen name, follower counts, description, retweet count, favorite count, friends count, status count,* etc. By using screen name as a unique identifier, we found that a total 8299 tweet users tweeted 3081 publications[2]. Table 1 shows the statistics of the dataset used in the experimentation.

There are a significant number of papers for which we find no tweets in our selected dataset. We decided to keep only those papers which have at least one tweet. Based on our cross-matching between references and tweet data set, we were left with 1537 papers which have a complete reference list and at least one tweet user interaction. The final dataset consists of 6272 references that were cited in at least one paper and 1551 twitter users that interact with at least one paper.

Table 1: Descriptive statistics of the Twitter dataset

Description	**Value**
Number of papers	3081
Unique Twitter User	8299
Publication time window	July 2011 to June 2016

[2] The data and code to reproduce or extend this work is available at the following URL: https://github.com/slab-itu/tweet_coupling

4.2 Evaluation measures

In order to evaluate our methodology, we compute the confusion matrix. The confusion matrix is given in Table 2. A confusion matrix contains four entries including (i) True Negative (TN); True Positive (TP); (iii) False Negative (FN); and (iv) False Positive (FP). In the context of bibliographic coupling and tweet coupling, we define these terms as follows (see Table 2).

When the publications are actually bibliographic- and tweet coupled (i.e., Actual "YES") and:

a) *True Positive (TP):* our methodology predicted "YES" (i.e., they are bibliographic- and tweet coupled);
b) *False Positive (FP):* our methodology predicted "NO" (i.e., they are *not* bibliographic-and tweet coupled).

When the publications are actually *not* bibliographic- and tweet coupled (i.e., Actual "NO") and:

c) *True Negative (TN):* our methodology predicted "NO" (i.e., they are *not* bibliographic-and tweet coupled);
d) *False Negative (FP):* our methodology predicted "YES" (i.e., they are bibliographic-and tweet coupled).

Once we obtained the confusion matrix, we evaluated the performance of our solution using the following seven evaluation measures which can be derived from confusion matrix.

i. *Accuracy:* The accuracy indicates that, overall how often our methodology predicts correctly (i.e., (TP+TN) / Total).
ii. *Misclassification Rate (*MR):* The *MR indicates that, how often our methodology is wrong (i.e., (FP+FN) / Total).
iii. *True Positive Rate (*TPR):* The *TPR indicates that, when the publications are actually bibliographic- and tweet coupled (i.e., yes), how often does our methodology predicts yes (i.e., TP / Actual Yes).

iv. *False Positive Rate (*FPR):* The *FPR indicates that, when the publications are actually *not* bibliographic- and tweet coupled (i.e., no), how often does our methodology predicts Yes (i.e., FP / Actual No).

v. *Specificity:* The specificity indicates that, when the publications are actually *not* bibliographic- and tweet coupled (i.e., no), how often does our methodology predicts no (i.e., TN / Actual No).

vi. *Precision:* The precision indicates that, when our classifier methodology yes, how often it is correct (i.e., TP / Predicted Yes).

vii. *Prevalence:* The prevalence indicates that how often does the yes condition actually occurs in our dataset (i.e., Actual Yes / Total).

4.3 Bibliographic and tweet coupling comparison

Table 2 shows the confusion matrix from which we obtain as a result of the Jaccard similarity between bibliographic coupling and tweet coupling. The total 0 elements in bibliographic coupling matrix are 2,346,508 where total of 0 elements in tweet coupling are 2,306,376. Count of nonzero items is respectively 15,861 and 55,993.

Table 3 shows the values of binary classifier from our confusion matrix. While the similarity results show high accuracy of 97%, we observe low True Positive Rate (TPR) and Precision. In order to further investigate the relation between bibliographic coupled and tweet coupled papers, we empirically apply different thresholds on a number of common twitter users and references.

Table 2: Bibliographic coupling and tweet coupling comparison confusion matrix

	Predicted (NO)	**Predicted (YES)**	**Total**
Actual (NO)	2,295,346 (TN)	51,162(FP)	2,346,508
Actual (YES)	11,030 (FN)	4,831(TP)	15,861
Total	2,306,376	55993	N= (2,362,369)

Table 3: Results of comparison between bibliographic coupling and tweet coupling.

Accuracy	***MR**	***TPR**	***FPR**	**Specificity**	**Precision**	**Prevalence**
97%	2.63%	30%	2.18%	97.8%	8%	0.67%

* MR= Misclassification rate; TPR = True Positive Rate; FPR = False Positive Rate

Table 4: Evaluation results of bibliographic coupling (BC) and tweet coupling (TC) and different thresholds

	Measures	**BC**	**TC**	**BC**	**TC**	**BC**	**TC**
		References **>= 10**	Twitter Users **>=10**	References **>= 5**	Twitter Users **>= 5**	References **>=3**	Twitter Users **>= 3**
Accuracy	(TP+TN)/total	94%	75%	96%	89%	96%	92%
MR	(FP+FN)/total	5%	24%	4.20%	10%	3.50%	7.90%
TPR	TP/actual yes	31%	93%	30%	86%	29%	74%
TNR	FP/actual no	3%	25%	3%	10%	2.50%	7.60%
Specificity	TN/actual no	96%	74%	97%	89%	97%	92%
Precision	TP/predicted yes	30%	16%	18%	14%	14%	11%
Prevalence	actual yes/total	3%	4%	2%	1.95%	1.44%	1.26%

Table 4 shows the evaluation results of bibliographic coupling and tweet coupling for different thresholds. With at least 10 common references and 10 common tweet users between papers, the reported accuracy is 94% and 75% for bibliographic coupling and tweet coupling, respectively. For this purpose, we set the threshold value to 5 for bibliographic coupling and tweet coupling. As shown in Table 4, the accuracy of bibliographic coupling and tweet coupling with the threshold value of 5 to increase 94% to 96% and 75% to 89% respectively. To maximize the value of accuracy and true positive rate in bibliographic coupling and tweet coupling, we set the threshold value for tweet coupling is more than 3 common twitter user tweets about the paper and for bibliographic coupling at least 3 common references in each paper. Our empirical evaluation suggests that the best similarity match between bibliographic coupled and tweet coupled is achieved at a threshold value of at least 3 references and 3 tweet users interaction per coupled paper. The accuracy of bibliographic coupling does not change but true positive rate drops to 1% and also other values Misclassification rate, False positive rate, and specificity not significantly change. On the other hand, as for tweet coupling, the Accuracy increased to 92% with true positive rate of 74%. Misclassification rate and false positive rate decreased to 7.9% and 7.6% respectively and specificity increased to 92%.

4.4 Bibliographic and Tweet coupling network comparison

Further, we create a network of bibliographic coupling matrix and tweet coupling matrix using VOSviewer software[3]. VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks (clusters) can be constructed based on bibliographic coupling, tweet coupling. From our bibliographic and tweet coupling matrix (see section 3.3 and 3.4), we visualise the relationship among papers in Fig. 2 and Fig .3, respectively. Note that each paper is represented with the source title (journal or conference they published in) concatenated with a system generated unique paper identification number. Using the dataset of 1537 publications that are both bibliographically and tweet coupled, the visualisation approach helps to understand how papers are clustered with respect to source titles.

Figure 2 shows the bibliographic coupling network grouped in 26 clusters. The maximum value of publications in a cluster is 141 and the minimum value is 2. Further, Figure 3 is a visualisation of tweet coupling network graph of publications, grouped in 17 clusters, where a maximum number of publications in a cluster are 201 and minimum in a cluster are 4 in numbers. Drilling down to these created networks further Figure 4, and Figure 5 demonstrate the clustering using bar graphs by bibliographic coupling and tweet coupling respectively. We removed all the journals from the cluster if they have < 4 papers in a cluster, then we are left with 22 clusters out of 26 in bibliographic coupling and with 16 out of 17 clusters in tweet coupling.

[3] http://www.vosviewer.com

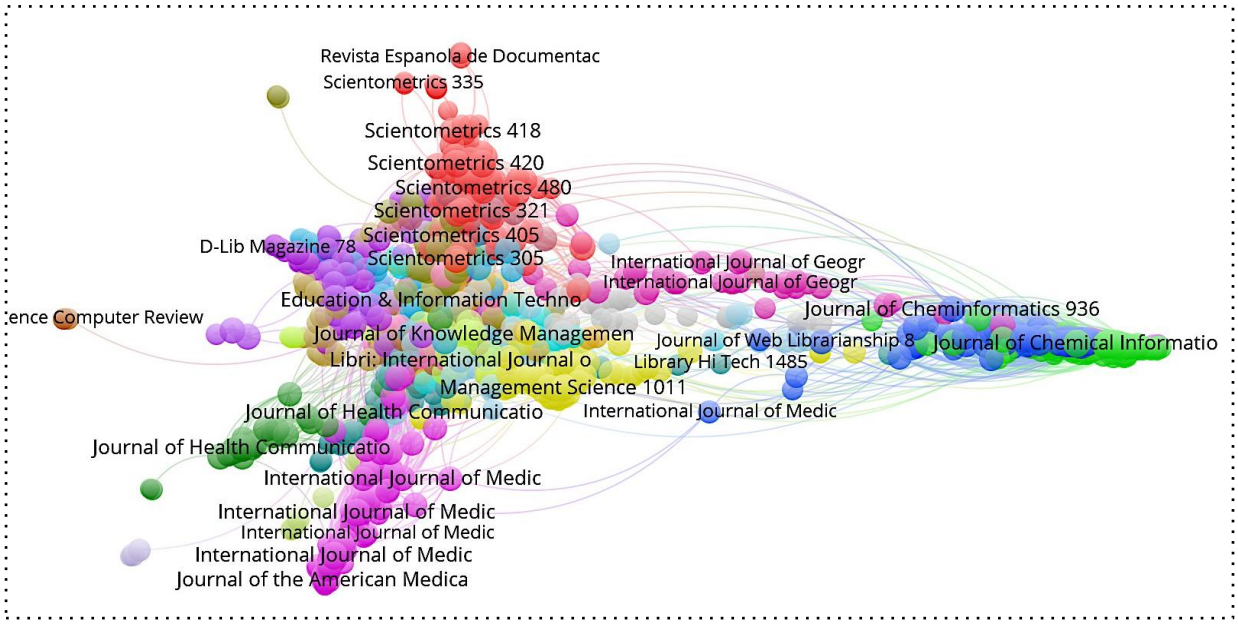

Figure 2: A Visualization of bibliographic coupling network of publications

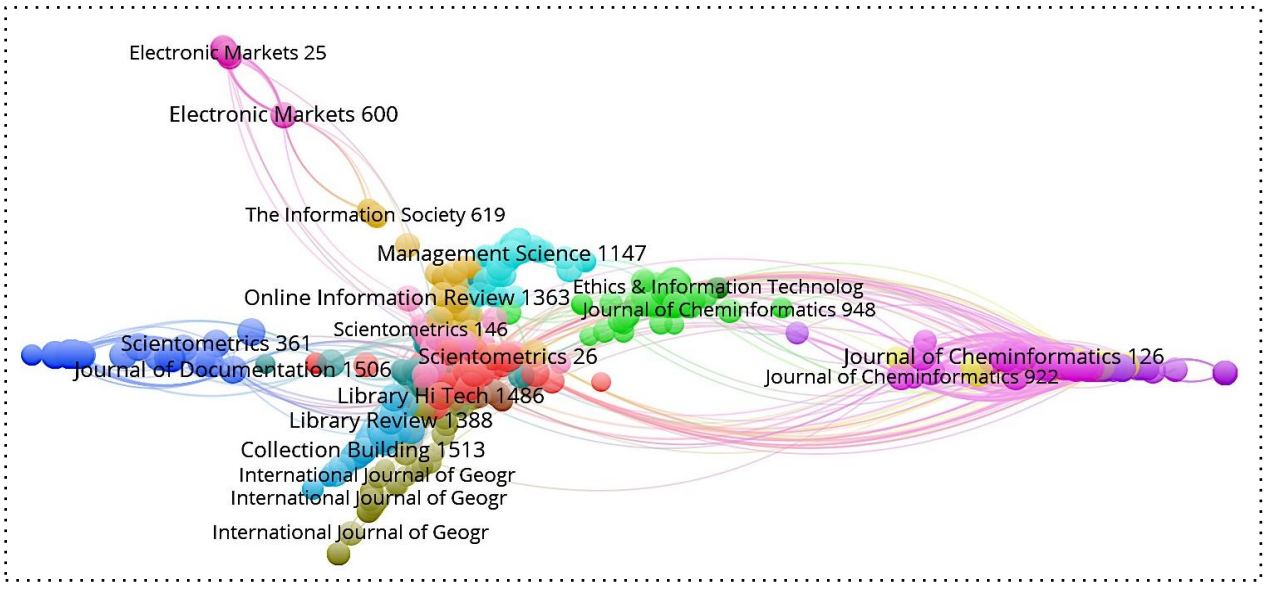

Figure 3: A visualization of tweet coupling network of Publications

The analysis shows a number of similar clusters, in terms of the presence of journals, both in using bibliographic coupling and tweet coupling, respectively: C-2, C-3, C-4, C-15, and C-23. In tweet coupling, scientific communities working in similar field are more connected as compare to bibliographic coupling. In contrast to tweet coupling, bibliographic coupling-based clusters show journals from different subfields within the Library and Information Sciences e.g. in bibliographic coupling “Collection Building” journal and “Journal of Health Communication” fall together in

cluster C-15, but in tweet coupling "Collection Building Journal" grouped with core journals of Library and Information Sciences in cluster C-7. Similarly, bibliographic based clustering shows "Journal of Health Communication" in cluster C-15, grouped with journals associated with core Library and Information Sciences journals, in contrast, tweet coupling based clustering shows the same journal grouped with other journals in the subfield of health informatics, in cluster C-2. We also see that using bibliographic based clustering, Electronic Markets Journal appears in C-8 and C-18 with the journals related to different subfields of Library Information Science, but in tweet coupling based clustering it appears in a single cluster.

Overall, the clustering results show that bibliographic coupling and tweet coupling based clustering complement each other in terms of grouping similar papers in a respective cluster. However, the tweet coupling based clustering highlights an interesting phenomenon i.e. the tweet user on social media networks are interested in similar subfields within Library and Information Sciences, in contrast to bibliographic based clustering, which groups cross-disciplinary journals within a cluster.

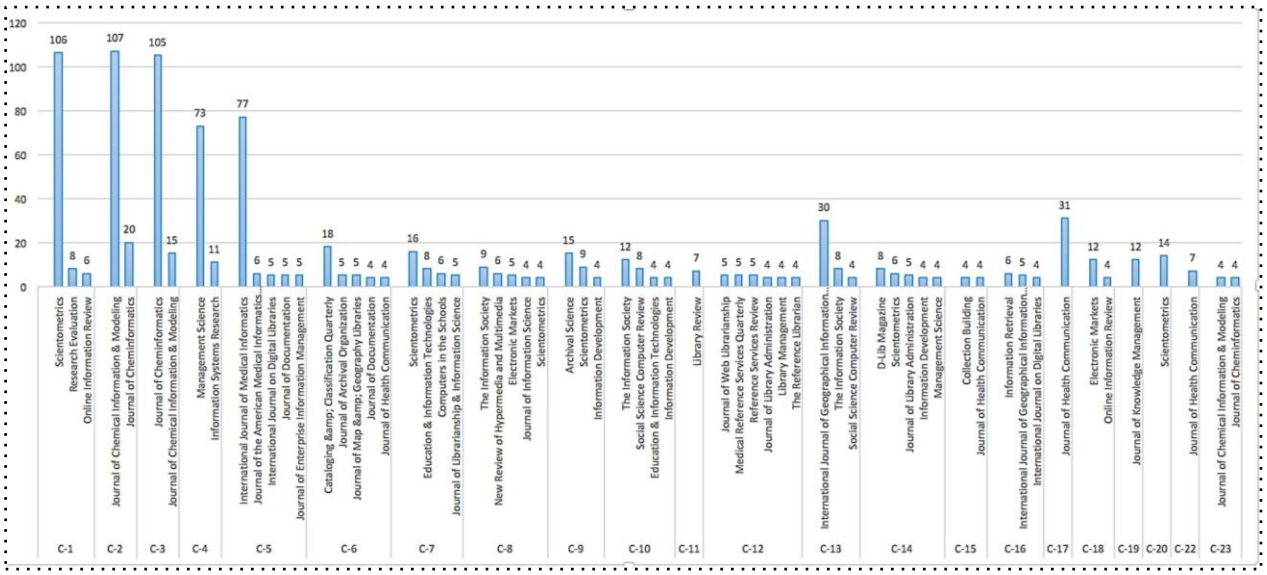

Figure 4: Result of Papers clustering by journals using bibliographic coupling

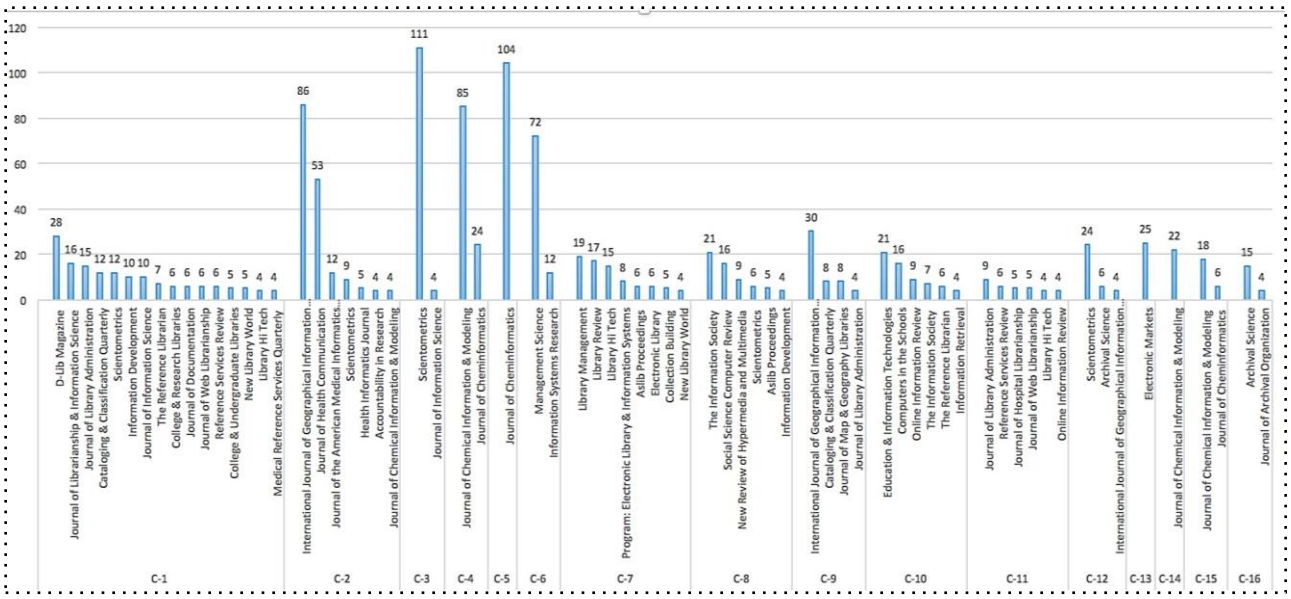

Figure 5: Result of papers clustering by journals using tweet coupling

5 Concluding remarks

In this study, we have examined the similarity of documents on behalf of their social usage by online communities on twitter platform and cited reference by the authors of the publications. We propose the concept of tweet coupling, which is a methodology for clustering scientific documents taking into account their social usage, whereas, we used a bibliographic coupling to find the similarity among the publications from the author's perspective. Our analysis shows that journals associated within a subfield strongly connected with each other in tweet coupling - whereas bibliographic based clustering shows cross-disciplinary journals within a group. We believe that tapping the advancements of crowdsourcing data provides a unique perspective of online social media community engaged with the scientific publications. More specifically, in contrast to conventional approaches like bibliographic coupling or co-citation that comes along with the inherent issue of publication and citation time lags, the tweet coupling has the ability to determine the similarity between papers based on real-time usage or discussion of scientific literature on social media platforms. Nevertheless, the phenomenon of tweet coupling is well suited since user perception is important to group publications for scholarly data management point of views such as clustering, classification or information retrieval. Also, in contrast to traditional bibliographic based approaches, the tweet coupling based method can group fine grained clustering down to sub-disciplines with a broader discipline for improved document management.

While reporting a significantly reliable accuracy, there are some limitations of this method. We found that not all the publications are discussed on Twitter, so a portion of publication dataset has to be discarded before the comparison can be performed. Prevalence of corrupted DOI's in altmetrics data set also hinder wider applications of this method. In the future, we plan to find similarity between publications by incorporating tweet text and document title and abstract text to compute tweet and bibliographic coupling metrics. We believe that by co-word analysis of tweets and papers title and abstract can produce an interesting result to figure out the semantic relation between social usage and bibliographic usage of references.

Further studies can also look for tweet sentiments such as positive, negative and natural, papers with higher positive sentiments tweets can be assigned higher weight while evaluating the research impact of publications which may improve the citation prediction results. It is possible that most recent publications have received more attention on social media as the usage of social media increased among scholars, but these publications may receive less citation count due to less time since published, therefore considering the time span while predicting the citation count may improve the result by considering tweet sentiments.

Specific to discipline, social usage helps us to determine the communication and writing style of the discipline. Semantic analysis of those tweets which belong to influential network nodes produces interesting results. Social network analysis can be used to establish a relationship between influential tweeters and relational structure of social media. Last but not the least, in our current approach only considered Twitter to find the relationship between social citation and academic citation, we can expand on this by including multiple social media platform like Facebook, Google+, etc. and potentially improve the results.

We believe that tweet coupling can further be exploited in future studies for the scientific document search applications such as classification of scientific documents, recommendation systems, and information retrieval systems for digital libraries.

Acknowledgments

The authors (Saeed-Ul Hassan & Mudassir Shabbir) were funded by the CIPL (National Center in Big Data and Cloud Computing (NCBC) grant, received from the Planning Commission of

Pakistan, through Higher Education Commission (HEC) of Pakistan. This work was partially supported by the Spanish Ministry of Science and Technology under the projects TIN2017-89517-P and TIN2017-83445-P. Eugenio Martínez Cámara was supported by the Spanish Government Programme Juan de la Cierva Incorporación (IJC2018-036092-I).

References

Adie, E., & Roe, W. (2013). Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, *26*(1), 11–17.

Amsler, R. A. (1972). *Applications of citation-based automatic classification*. Linguistics Research Center, University of Texas at Austin.

Ananiadou, S., Thompson, P., & Nawaz, R. (2013). Enhancing search: Events and their discourse context. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 318-334). Springer, Berlin, Heidelberg.

Batista-Navarro, R. T., Kontonatsios, G., Mihăilă, C., Thompson, P., Rak, R., Nawaz, R., ... & Ananiadou, S. (2013). Facilitating the analysis of discourse phenomena in an interoperable NLP platform. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 559-571). Springer, Berlin, Heidelberg.

Butler, J. S., Kaye, I. D., Sebastian, A. S., Wagner, S. C., Morrissey, P. B., Schroeder, G. D., … Vaccaro, A. R. (2017). The evolution of current research impact metrics: from bibliometrics to altmetrics? *Clinical Spine Surgery*, *30*(5), 226–228.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019.

Didegah, F., & Thelwall, M. (2018). Co-saved, co-tweeted, and co-cited networks. Journal of the Association for Information Science and Technology, 69(8), 959-973.

Erdt, M., Nagarajan, A., Sin, S.-C. J., & Theng, Y.-L. (2016). Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, *109*(2), 1117–1166.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, *1*(4), 359–375.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Jama*, *295*(1), 90–93.

Gipp, B., & Beel, J. (2009). Citation proximity analysis (CPA): A new approach for identifying related work based on co-citation analysis. *ISSI'09: 12th International Conference on Scientometrics and Informetrics*, 571–575.

Habib, R., & Afzal, M. T. (2019). Sections-based bibliographic coupling for research paper recommendation. Scientometrics, 1-14 (in press).

Hassan, S.-U., & Gillani, U. A. (2016). Altmetrics of" altmetrics" using Google Scholar, Twitter, Mendeley, Facebook, Google-plus, CiteULike, Blogs and Wiki. *ArXiv Preprint ArXiv:1603.07992*.

Hassan, S. U., & Haddawy, P. (2013). Measuring international knowledge flows and scholarly impact of scientific research. Scientometrics, 94(1), 163-179.

Hassan, S. U., & Haddawy, P. (2015). Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy. Scientometrics, 103(1), 33-46.

Chicago

Hassan, S.-U., Imran, M., Gillani, U., Aljohani, N. R., Bowman, T. D., & Didegah, F. (2017). Measuring social media activity of scientific literature: an exhaustive comparison of scopus and novel altmetrics big data. *Scientometrics*, *113*(2), 1037–1057.

Haustein, S., Bowman, T. D., & Costas, R. (2015). Interpreting" altmetrics": viewing acts on social media through the lens of citation and social theories. *ArXiv Preprint ArXiv:1502.05701*.

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PloS One*, *10*(3), e0120495.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated “bot” accounts on T witter. *Journal of the Association for Information Science and Technology*, *67*(1), 232-238.

Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. Scientometrics, 116(2), 1367-1382.

Hellsten, I., & Leydesdorff, L. (2017). Automated Analysis of Topic-Actor Networks on Twitter: New approach to the analysis of socio-semantic networks. *ArXiv Preprint ArXiv:1711.08387*

Hellsten, I., Opthof, T., & Leydesdorff, L. (2019). N-mode network approach for socio-semantic analysis of scientific publications. *Poetics*, 101427.

Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, *101*(2), 1027–1042.

Huang, A. (2008). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).

Joubert, M., & Costas, R. (2019). Getting to Know Science Tweeters: A Pilot Analysis of South African Twitter Users Tweeting About Research Articles. *Journal of Altmetrics*, *2*(1).

Karimi, S., Moraes, L., Das, A., Shakery, A., & Verma, R. (2018). Citance-based retrieval and summarization using IR and machine learning. Scientometrics, 116(2), 1331-1366.

Katz, J. A. (2003). The chronology and intellectual trajectory of American entrepreneurship education: 1876–1999. *Journal of Business Venturing*, *18*(2), 283–300.

Kaufmann, A., & Kasztler, A. (2009). Differences in publication and dissemination practices between disciplinary and transdisciplinary science and the consequences for research evaluation. *Science and Public Policy*, *36*(3), 215–227.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.

Lawrence, S., Bollacker, K., & Giles, C. L. (1999). Indexing and retrieval of scientific literature. *Proceedings of the Eighth International Conference on Information and Knowledge Management*, 139–146. ACM.

Liu, X. Z., & Fang, H. (2017). What we can learn from tweets linking to research papers. *Scientometrics*, *111*(1), 349–369.

Lowry, P. B., Humpherys, S. L., Malwitz, J., & Nix, J. (2007). A scientometric study of the perceived quality of business and technical communication journals. *IEEE Transactions on Professional Communication*, *50*(4), 352–378.

Martyn, J. (1964). Bibliographic coupling. Journal of documentation, 20(4), 236-236.

Mesbah, S., Fragkeskos, K., Lofi, C., Bozzon, A., & Houben, G. J. (2017). Facet embeddings for explorative analytics in digital libraries. In *International Conference on Theory and Practice of Digital Libraries* (pp. 86-99). Springer, Cham.

Melero, R. (2015). Altmetrics–a complement to conventional metrics. *Biochemia Medica: Biochemia Medica*, *25*(2), 152–160.

Moed, H. F. (2011). The source normalized impact per paper is a valid and sophisticated indicator of journal citation impact. *Journal of the American Society for Information Science and Technology*, *62*(1), 211–213.

Mohammadi, E., & Thelwall, M. (2014). M endeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, 65(8), 1627-1638.

Nawaz, R., Thompson, P., & Ananiadou, S. (2012). Identification of Manner in Bio-Events. In *LREC* (pp. 3505-3510).

Peoples, B. K., Midway, S. R., Sackett, D., Lynch, A., & Cooney, P. B. (2016). Twitter predicts citation rates of ecological research. *PloS One*, *11*(11), e0166570.

Priem, J., & Costello, K. L. (2010). How and why scholars cite on Twitter. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *ArXiv Preprint ArXiv:1203.4745*.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*.

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: exploring the insides of Altmetric. com. ArXiv Preprint ArXiv:1408.0135.

Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: identifying main topics and actors. *FEMS microbiology letters*, *366*(7), fnz075.

Robinson-García, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PloS one*, *12*(8).

Safder, I., & Hassan, S. U. (2019). Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics*, 119 (1), 257- 277.

Said, A., Bowman, T. D., Abbasi, R. A., Aljohani, N. R., Hassan, S. U., & Nawaz, R. (2019). Mining network-level properties of Twitter altmetrics data. *Scientometrics*, 120 (1), 217–235.

Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. BMC medical informatics and decision making, 18(1), 46.

Shu, F., Lou, W., & Haustein, S. (2018). Can Twitter increase the visibility of Chinese publications? *Scientometrics*, *116*(1), 505–519.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

Sternitzke, C., & Bergmann, I. (2009). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, *78*(1), 113–130.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PloS One*, *8*(5), e64841.

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting links to academic articles. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, (17), 1–8.

Thijs, B., & Glänzel, W. (2018). The contribution of the lexical component in hybrid clustering, the case of four decades of "Scientometrics". Scientometrics, 115(1), 21-33.

Trueger, N. S., Thoma, B., Hsu, C. H., Sullivan, D., Peters, L., & Lin, M. (2015). *The altmetric score: a new measure for article-level dissemination and impact*.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.

Zahedi, Z., Costas, R., Larivière, V., & Haustein, S. (2017). What makes papers visible on social media? An analysis of various document characteristics. *ArXiv Preprint ArXiv:1703.05777*.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, *101*(2), 1491–1513.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology*, *65*(5), 995–1006.